Spring 5-7-2021

Parameter Estimation in Nonlinear Regression: Exploring Confidence Intervals for Estimated Coefficients

Brandon Williams

Parameter Estimation in Nonlinear Regression: Exploring Confidence Intervals for Estimated Coefficients

Brandon M. Williams

submitted in partial fulfillment of the requirements for Honors in Mathematics at the University of Mary Washington

Fredericksburg, Virginia

April 2021

This thesis by **Brandon M. Williams** is accepted in its present form as satisfying the thesis requirement for Honors in Mathematics.

DATE	APPROVED
5/6/2021	Debra L Hydorn Debra. Hydorn, Ph.D. thesis advisor
05/06/2021	Julius Esunge Julius. Esunge, Ph.D. committee member
05/05/2021	MDenhere Melody. Denhere, Ph.D. committee member

Contents

Abstract

Previously, we explored generating data using four different curved source functions with normally distributed errors and then fitting various curved models to this generated data. The goal of this research was to study the reliability of the goodness-of-fit measures, such as R^2 and AIC, when the data originates from a curved source model. Ultimately, we found that, for logarithmic source models, AIC picks the correct source model most often, while R^2 selects either the logarithmic or quadratic source models. For quadratic source models, AIC picks the power source model most often, whereas R^2 selects the correct source model. Lastly, we expanded our project to investigate how AIC and R^2 perform when violating the assumptions for regression analysis and concluded that the presence of an influential point impacts goodness-of-fit measures more for logarithmic than quadratic source models.

This semester, we have explored producing bootstrap confidence intervals for the coefficients of nonlinear regression models to fit generated data for a variety of different curved models. More specifically, using R, I use a command for nonlinear modeling to find the best-fit curved model given data from a known curved association and then apply bootstrap methods to produce the confidence interval estimates. The datasets will be generated for models with different levels of curvature, amounts of variation in the points around the best-fit model for normal and non-normal error distributions, and with varying numbers of data points. The observed proportion of intervals containing the true parameter values are observed.

1 Introduction

Many introductory statistics courses cover the topic of simple regression because of the model's increasing use. A regression analysis is the process of estimating the relationship between an independent variable(s) and a dependent variable. Regression analysis has become pervasive for its ability to describe and form relationships between categorical and quantitative variables. More importantly, the technique provides an estimate for the strength and direction of the linear relationship between two or more variables. The most common regression analysis is a linear regression; although, there are other types of regressions such as Non-linear and Bayesian linear regressions. This statistical technique is primarily used for prediction and forecasting as well as inferring causal relationships between the independent and dependent variable(s). Usually, when researchers apply these techniques, they must answer the following question: Which factors have influence on the dependent variable? On the contrary, which factors can be ignored? Once the factors have been decided, the researchers must answer, how do those factors interact with each other? Furthermore, the researchers should comment on the validity of the factors. Nonetheless, regression analysis has become one of the more important statistical techniques due to the desired properties as described above; however, there are assumptions that must be checked in order to perform the technique correctly.

For each type of regression, before the analysis can be performed, there are basic assumptions that must be regarded to determine if these assumptions are violated. If these assumptions are not checked, the pitfalls may range from incorrect interpretation and prediction to biased coefficients and standard errors. It is also important to note that correlation is not causation and regression analysis is very sensitive to "bad" data. The sensitivity can be measured by different types of curvatures in each datum. Given the widespread use of this technique, there is the possibility of performing these regression analyses incorrectly, such as using a non-linear estimator with linear data. This incorrect fitting leads to the possibility of improper interpretation/prediction, biased coefficients, and standard errors.

1.1 Prior Research

In the Summer of 2020, during the Summer Science Institute, our research explored generating data using four different curved source functions with normally distributed errors and then fitting various curved models to this generated data. The goal of this research was to study the reliability of the goodness-of-fit measures, such as R^2 and AIC, when the data originates from a curved source model. The intention of the research attempts to answer two research questions: if two models fit equally well, does it matter which model researchers use? Furthermore, if researchers use R^2 or AIC, how often do they obtain the correct underlying model? The coefficient of determination, R^2, can be defined as:

$$R^2 = 1 - \frac{RSS}{TSS} \tag{1}$$

where RSS is the sum of squares of residuals squared and TSS is the total sum of squares. In essences, this goodness of fit measures the proportion of the variance in the dependent variable that is explained by the association with independent variable(s). Furthermore, it provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. The measure can be interpreted as the percentage of variation in Y explained by the model. Lastly, a larger R^2 indicates a better or best fit. Next, the Akaike Information Criterion, AIC, is defined as:

$$AIC = 2K - ln(\hat{L}) \tag{2}$$

where K is the number of estimated parameters in the model and $\hat{L}$ is the observed value of the maximum likelihood function for the model. When using the AIC as a goodness of fit measure, the general rule of thumb is that a lower value indicates a better or best fit. In essence, the AIC is a model selection measure as the AIC estimates the quality of each model, relative to each of the other models. In other words, the AIC tells us how much information is lost if we replace the data with the model.

Ultimately, the research concludes that for logarithmic source models, AIC is lowest for the correct source model most often, while R^2 is highest for either the logarithmic or quadratic source models. For quadratic source models, AIC is lowest for the power source model most often, whereas R^2 is highest for the correct source model. Lastly, the project was expanded to investigate how AIC and R^2 perform when violating the assumptions for regression analysis and concluded that the presence of an influential point impacts goodness-of-fit measures more for logarithmic than quadratic source models.

This research does not to address the question of the reliability of the estimated regression coefficients for each model and association. Thus, the main research question of this simulated study is to answer if a researcher has identified the correct underlying model, how often do the confidence intervals contain the true values of the model parameter? To answer the research question, this paper explores producing confidence intervals for the coefficients of nonlinear regression models to fit generated data for a variety of different curved models. Using R, this paper uses a command for nonlinear modeling to find the best-fit curved model given data from a known curved association and then applies bootstrap methods to produce the confidence interval estimates. The data sets are generated for models with different levels of curvature, amounts of variation in the points around the best-fit model for normal and non-normal error distributions, and with varying numbers of data points. Once the estimates are collected, then this paper uses confidence intervals to determine if the true value is contained in the interval. If there is a skew in the data and histogram, then this paper proceeds by using bootstrap confidence intervals to compare across different models. In addition, the observed proportion of intervals containing the true parameter values are observed and compared to the bootstrap confidence intervals.

1.1.1 Literature Review

The literature review surrounding the reliability of estimated coefficients for nonlinear regressions is scant. However, Samart et. all (2018) focuses on using the exact bootstrap to construct confidence intervals for regression parameters in small samples. Small sample sizes can violate the assumptions for regression analysis. The authors investigate the reliability of estimated coefficients when small sample sizes are used. To that end, the authors use bootstrapping methods to compare different confidence intervals. By uses a simulated study, the authors find that on a very small sample (n $\approx$ 5) under Laplace distribution with the independent variable treated as random, the exact bootstrap was more effective than the standard bootstrap confidence interval.

Although the literature review surrounding the reliability of estimated coefficients for nonlinear regressions is very limited, this paper aims to provide more in-depth analysis of examining parameter estimation for nonlinear regressions. Specifically, this paper contributes to the field by analyzing the reliability of estimated coefficients for Exponential, Power, Logarithmic, and Quadratic associations. To the knowledge of the author, there does not appear to be another simulation study similar to the one conducted.

For this research, there are four different nonlinear associations: Exponential, Power, Logarithmic, and Quadratic. The functional form for each of the model is as shown below:

$$y = a{\color{red}b}^x \tag{3}$$

$$y = ax^{\color{red}b} \tag{4}$$

$$y = a + {\color{red}b} * \ln(x) \tag{5}$$

$$y = a + bx + {\color{red}c}x^2 \tag{6}$$

where the estimated coefficient is shown in red. It is important to note that when the regression analysis is conducted, the research involves producing estimates and confidence intervals for the parameters marked in red.

The assumptions of the model do not need to be checked as well. Because we are using a single parameter, multicollinearity cannot occur whereas serial correlation cannot also occur as there is no time element to the data. Lastly, heteroskedasticity cannot occur either as the R code produces the error terms, which are normalized. Thus, we can proceed without worrying about violating the assumptions of the linear and nonlinear regression.

The paper will proceed as follows: Section II further discusses the methods of the project, Section III interprets the results, and finally, Section IV concludes.

2 Methods

Since the goal of the paper is to determine the reliability of estimated coefficients, the data can be generated by the researcher. This generated data gives the research the ability to create "good" and "bad" data. In this project, the data generated varies on three factors: number of observations, the amount of variation, and degree of curvature. The number of observations varies from 19, 37, and 91, whereas the amount of variation changes from 10%, 20% and 30%. Each of these values serves as a proxy for low, medium, and high levels. Specifically, when generating the data, the R code uses a command to generate a sequence from 1 to 10 with an interval width of 0.5 for 19, 0.25 for 37, and 0.1 for 91 observations. The sample sizes are obtained from the intervals width rather than a specific sample size.

2.1 Kappa - Degree of Curvature

The degree of curvature is calculated by using the parameter called kappa. Kappa is defined as

$$k = \frac{|f''(x)|}{(1+(f'(x))^2)^{\frac{3}{2}}} = \frac{1}{Radius} \tag{7}$$

It is important to note that the curvature is different at each point along the line of the function; however, the kappa value is determined by the level of curvature which maximizes the curvature of each function. For example, consider the example of the logarithmic function. Using the equation above, then the logarithmic function is as defined below:

$$f(x) = y = a + b * ln(x) \tag{8}$$

Thus, the first derivative and the second derivative are needed to calculate kappa:

$$\begin{aligned} f'(x) &= \frac{b}{x} \\ f''(x) &= \frac{(-b)}{x^2} \end{aligned} \tag{9}$$

Now, simply plug in the derivative to the formula of kappa. It is important to note that kappa is also equal to one over the radius.

$$\begin{aligned} k &= \frac{|\frac{b}{x^2}|}{(1+(\frac{b}{x})^2)^{\frac{3}{2}}} \\ f'(k) &= \frac{b}{\sqrt{2}} \end{aligned} \tag{10}$$

Below is a depiction of kappa for the logarithmic function. This visualization serves to demonstrate kappa. The blue line refers to the logarithmic function of y = 2 + 3*log(20). The red line is a construction of the circle tangent to the function. As can be seen, kappa is equal to one over the radius of the circle.

Visualization of Kappa for Logarithmic Function

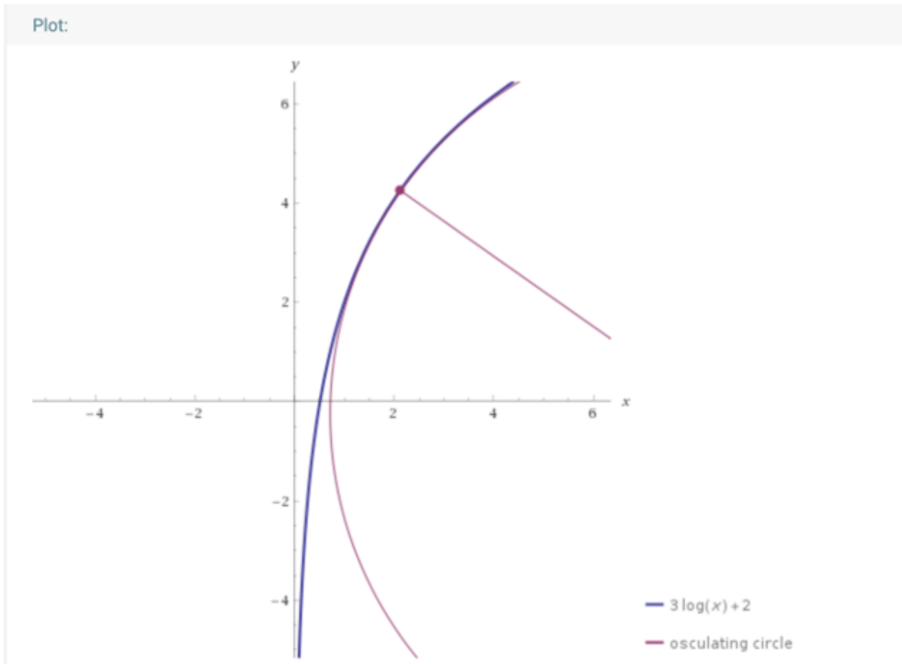

Figure 1: This graph shows the visualization of kappa. The blue line refers to the logarithmic function, whereas the red line refers to the circle which shows the point tangent to the function. Kappa is defined as one over the radius of that circle. Thus, it has been established how kappa can be seen.

To understand the different kappa values for each model, Figure 2 shows the different degree of curvatures for the Quadratic and the Logarithmic functions. It is evident that the degree of curvature varies greatly for each model.

Visualization of Kappa for Different Associations

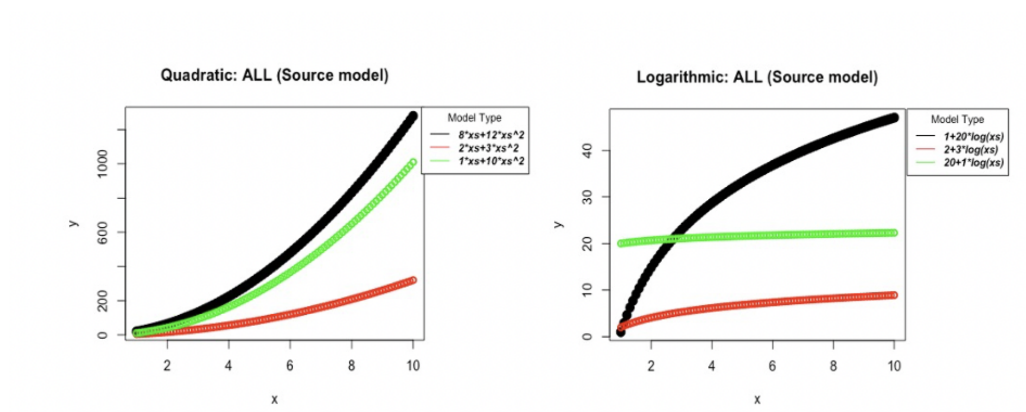

Figure 2: These graphs shows kappa for different associations, in particular, Quadratic and Logarithmic. This visualization is important to show the different degree of curvatures for each model and the incompatibility to interpolate from different models.

In general, kappa can be different for each model, thus the kappa for a specific association cannot be compared to a different association. It is important to note that this research also uses a proxy for low, medium, and high degrees of curvature; however, the values differ from model to model. Table 1 depicts these values.

Table of Kappa Values for Different Nonlinear Associations

	Exponential	*Power*	*Logarithmic*	*Quadratic*
Low	8*12^xs	1*xs^2	1+20*log(xs)	8*xs+12*xs^2
Medium	2*3^xs	2*xs^3	2+3*log(xs)	2*xs+3*xs^2
High	1*2^xs	8*xs^12	20+1*log(xs)	1*xs+10*xs^2

Figure 3: Table of Kappa values for each nonlinear association.

From low to high, the true value for logarithmic association are 20, 3, and 1 whereas for the quadratic association are 12, 3, and 10. Furthermore, for the exponential association, the true values are 12, 3, and 2 while for the power association are 2, 3, and 12. It is important to remember that this paper continues by stating the true value; however, the true value differs for each association and degree of curvature.

2.2 Creating an R program

After the data is generated, to test the reliability of the coefficients of regression models, we constructed a R code to generate 1000 random data sets for a specified model, number of points, amount of variation, and degree of curvature. For the Logarithmic function, the regressions estimates technique is the ordinary least squares (OLS), whereas for Exponential, Power, and Quadratic, the estimate technique is nonlinear least squares (NLS). In fact, in order to run the code efficiently and consistently, the R code uses the nlsLM command. This command is a modified version of nonlinear least squares; however, this command uses the method of Levenberg-Marquardt algorithm instead of the default method of Gauss-Newton algorithm. the It is important to note that the technique for Quadratic is slightly different as the code required three points to estimate the equation, whereas Exponential and Power required two points. This additional point led to more accurate fit compared to the other models. Once the regression has been conducted, the code captures the estimated coefficient, and lower bound and upper bound of the confident interval.

After the values have been stored, the code proceeds to measure if the histogram of the estimated coefficients are unimodal and symmetric for each model. To determine if there is sufficient skewness in the histogram, the code uses the following rule to assess the skew: a skewness greater than 0.5 is considered heavily skewed. If so, the code then uses bootstrapping techniques to transform the estimated coefficients. Lastly, the code creates histograms and regression plots to visual the entire process.

Once the estimates have been gathered, the paper determines if the 95% confidence interval contains the true value of the selected model parameters. Also, the R code collects the mean estimated coefficient for each nonlinear association. These three different measures are displayed in bar charts to measure the difference in number of observations, amount of variation, and degree of curvature.

3 Results

Before exploring the results, it is important to discuss the a priori assumptions about the results. In general, it is expected that each of the non-linear models have the true value contained in the confidence interval about 95% of the time. In addition, for the mean of the estimate coefficients, it is expected that the estimate be near their true value; however, variation is a factor as the R code is programmed to give the y-values a certain variation for each trial. The exact values, which are used to create each of the plots, can be found in the Appendix.

3.1 True Value within the Confidence Interval

As mentioned earlier, the R code produces a 95% confidence interval for each simulation. Afterwards, the code proceeds to check if the true value is contained within the confidence interval. It is important to note that if the true value is contained, then this result is classified as a success or 1, thus a 92% means that for 1000 trials, there is 920 times in which the true value is contained. The results for the True Value within the Confidence Interval are depicted in the table below. As can be seen, there is a graph for each degree of curvature. Within each graph, there is a trial for each number of observations and then each amount of variation.

Low Degree of Curvature

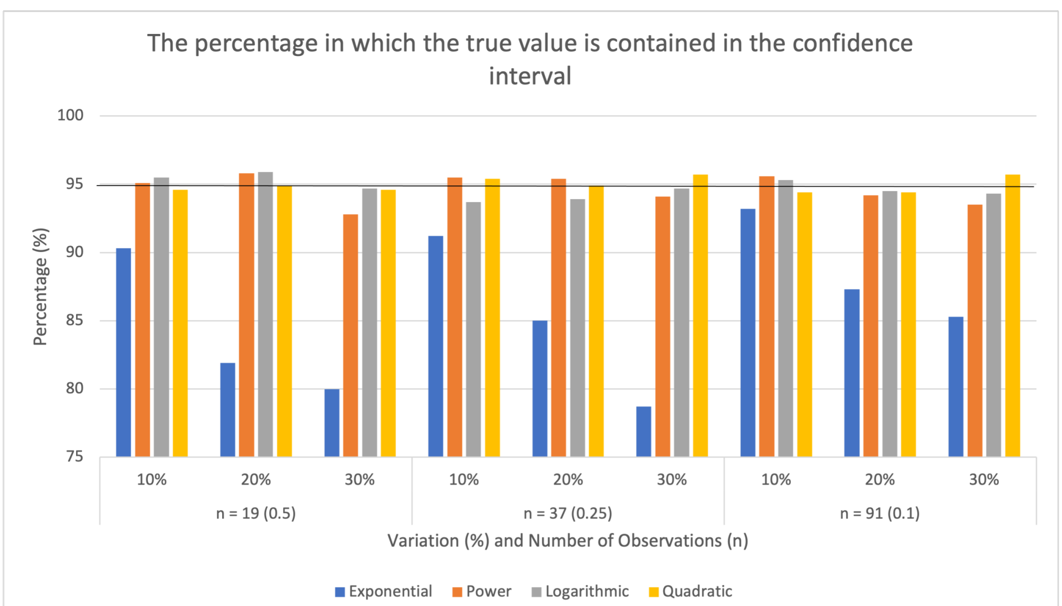

Figure 4: Results for the low degree of curvature. Within the plot, there is a trial for each number of observations and amount of variation. In total, there is 9 different trials.

For the low degree of curvature, the Logarithmic, Power, and Quadratic performances seem to be very consistent and reliable as the true value is contained within the confidence interval at least or approximately 95% of the time. On the contrary, the Exponential performance is very sensitive to the amount of variation. It is important to note that as the number of observations increases, the results for the Exponential performance seem to decrease the capture percentage less as the amount of variation increases.

Medium Degree of Curvature

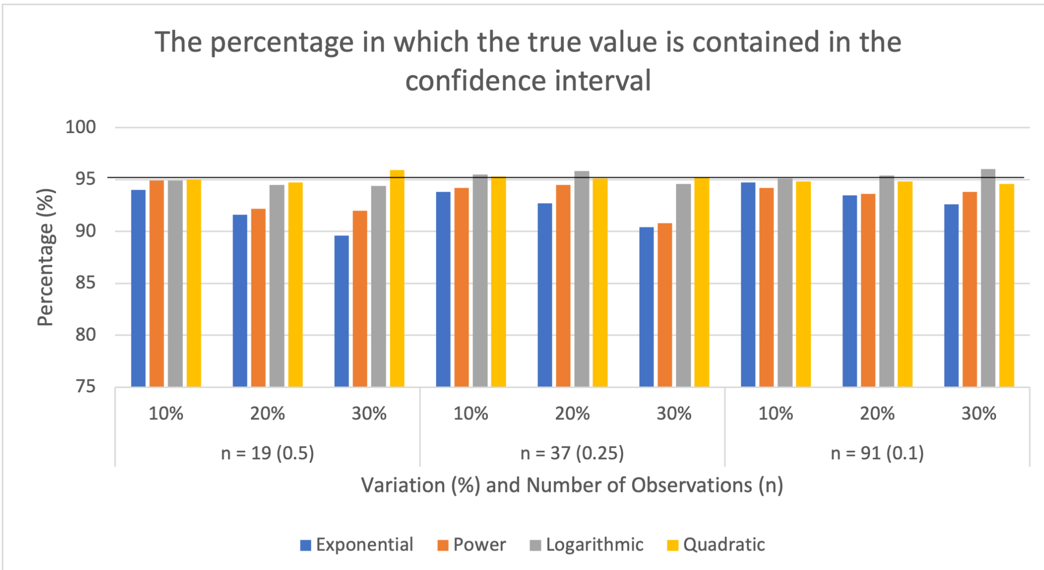

Figure 5: Results for the medium degree of curvature. Within the plot, there is a trial for each number of observations and amount of variation. In total, there is 9 different trials.

For the medium degree of curvature, Exponential, Power, and Logarithmic performances are sensitive to changes in the amount of variation. The Quadratic performance seems to be the most consistent and reliable among the performances. Although, it seems that as the number of observations increases, the reliability increases because the percentage in which the true value is contained with in the confidence interval increases. As the amount of variation increases, the percentage tends to decrease.

High Degree of Curvature

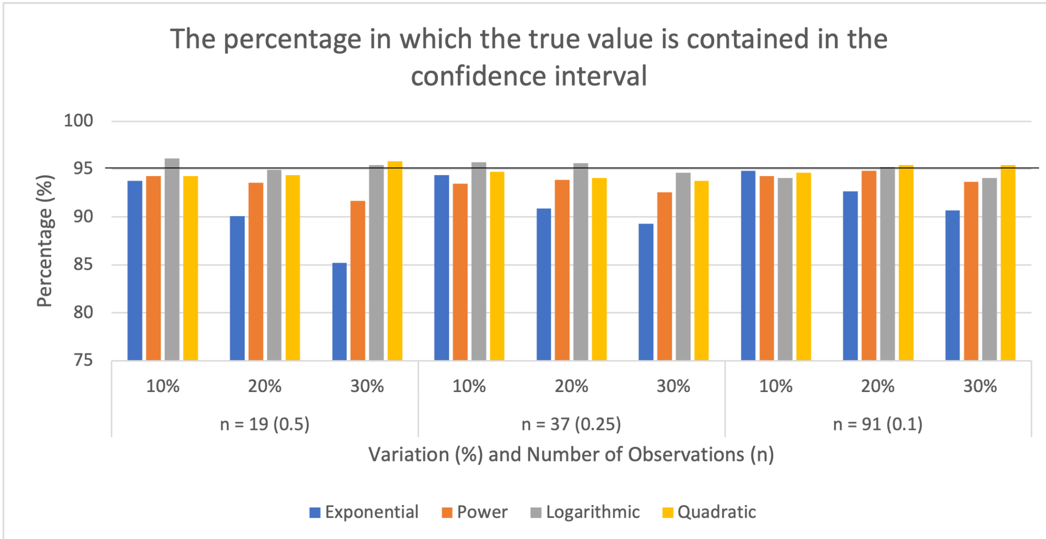

Figure 6: Results for the high degree of curvature. Within the plot, there is a trial for each number of observations and amount of variation. In total, there is 9 different trials.

For the high degree of curvature, it can be seen that Exponential, Power, and Logarithmic performances are sensitive to changes in the variation. The Quadratic performance seems to be the most consistent and reliable among the performances. Although, it seems that as the number of observations increases, the reliability increases because the percentage in which the true value is contained with in the confidence interval increases. As the amount of variation increases, the capture percentage tends to decrease.

When comparing each of the degrees of curvature, it appears that increasing the degree of curvature leads to higher percentages in which the true value is contained in the confidence interval. It is believed that this result arises because the associations are nonlinear and as the data generated becomes more nonlinear (i.e. more curve to the data), the regression techniques, which use nonlinear estimations for Exponential, Power, and Quadratic, are able to generate better curves of best fit and more reliable estimated coefficients that are close to the true value.

3.2 Mean of the Estimate Coefficient

The results for the mean estimate coefficient are below. For the low degree of curvature, the true value for each association is as follows: 12 for Exponential, 3 for Power, 20 for Logarithmic, and 12 for Quadratic. As can be seen, the Power, Logarithmic, and Quadratic are very reliable in their mean of the estimate coefficient being exactly the true value. On the contrary, the Exponential association seems to be unreliable as the mean changes drastically. In fact, for a large sample size and 10% of variation, the mean of the estimated coefficient is the closest to the true value but not exact compared to the other trials.

Low Degree of Curvature

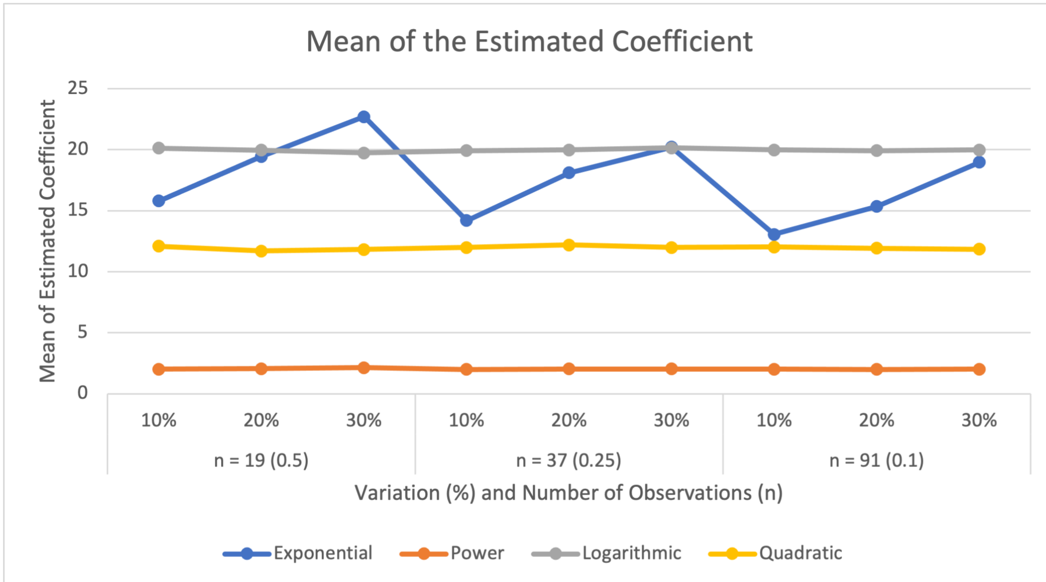

Figure 7: Results for the low degree of curvature. Within the plot, there is a trial for each number of observations and amount of variation. In total, there is 9 different trials.

For the medium degree of curvature, the true value for each association 3. It is evident that the Exponential association is not reliable as the mean value is not close to the true value. Although, it is important to note that when the variation is small, the mean value is approximately the true value, an improvement from the low degree of curvature where it is only true for the large sample size. Furthermore, the Power association seems to be a little less reliable; however, overall, the mean estimated coefficient is approximately the true value. Lastly, the Logarithmic and Quadratic associations are reliable and consistent.

Medium Degree of Curvature

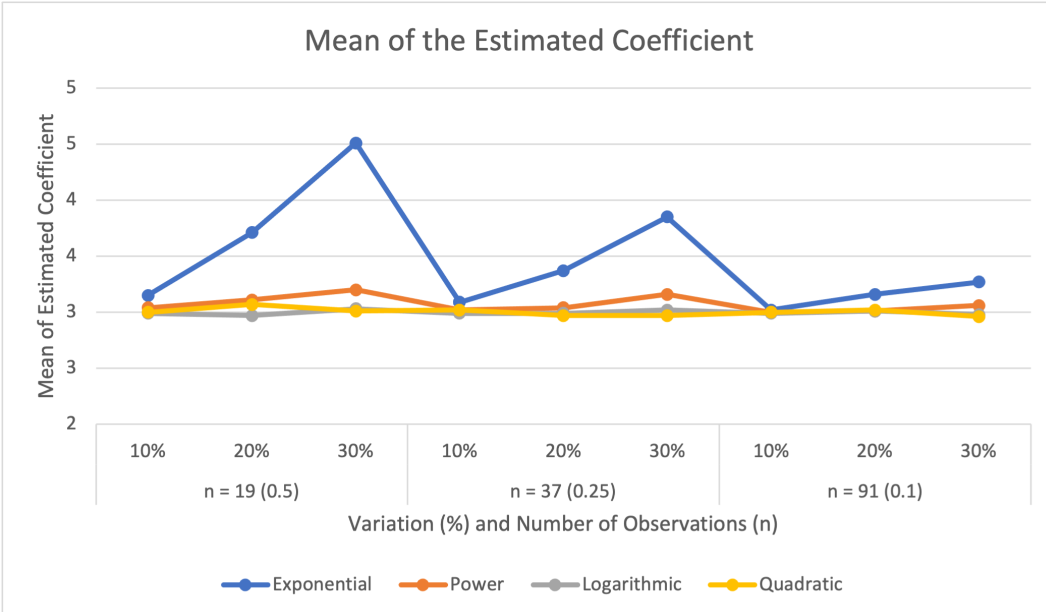

Figure 8: Results for the medium degree of curvature. Within the plot, there is a trial for each number of observations and amount of variation. In total, there is 9 different trials.

For the high degree of curvature, the true value for each association is as follows: 2 for Exponential, 12 for Power, 1 for Logarithmic, and 10 for Quadratic. It seems that each association is fairly reliable and consistent. There are trials where the mean value is not exactly the true value; however, overall, each of the associations are fairly reliable.

High Degree of Curvature

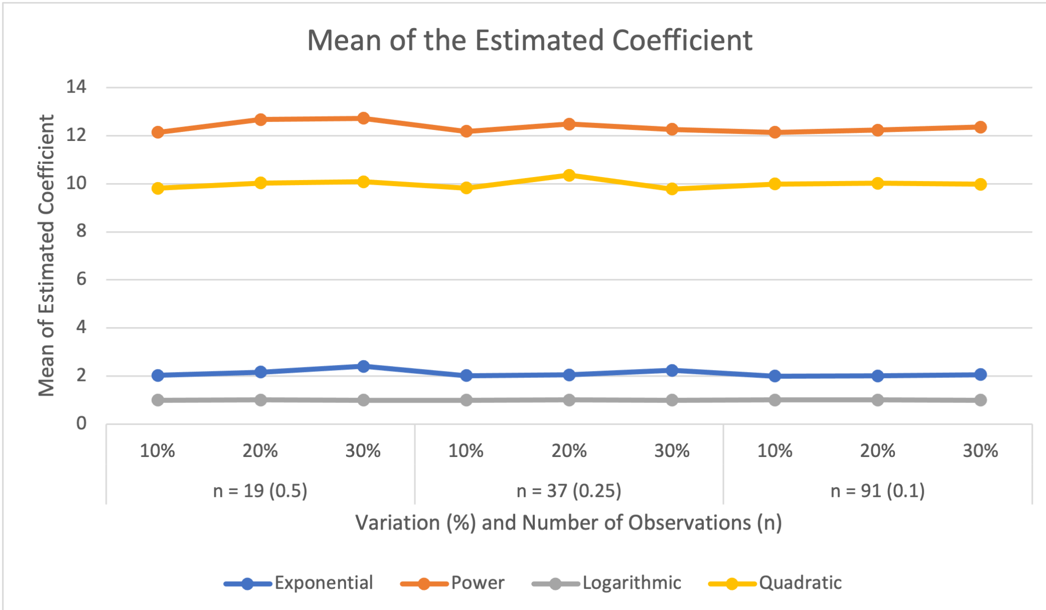

Figure 9: Results for the high degree of curvature. Within the plot, there is a trial for each number of observations and amount of variation. In total, there is 9 different trials.

When comparing the degree of curvature, it appears that increasing the degree of curvature leads to the mean value of the estimated coefficients being closer to the true value of the association. It seems that the Power, Logarithmic, and Quadratic associations are fairly consistent regardless of the degree of curvature; however, the Exponential association appears to become more reliable as the degree of curvature decreases.

4 Conclusion

This paper strives to explored producing confidence intervals for the coefficients of nonlinear regression models to fit generated data for a variety of different curved models. More specifically, using R, this paper make use of a command for nonlinear modeling to find the best-fit curved model given data from a known curved association and then produce the confidence interval estimates. If the histograms of the estimated coefficients are not unimodal and symmetric, then this paper applies bootstrap methods to produce bootstrap confidence intervals. The datasets will be generated for models with different levels of curvature, amounts of variation in the points around the best-fit model for normal and non-normal error distributions, and with varying numbers of data points. The observed proportion of intervals containing the true parameter values are observed.

This paper finds that a higher degree of curvature leads to an increase in the reliability and consistency of the estimates coefficients. Furthermore, the number of observations also increases the reliability and consistency of the estimates coefficients. Finally, the amount of variations plays an important role as the greater the variation in the data, the estimate coefficients are less reliable.

Researchers must strive to maximize their sample size; however, they should be cautious about the amount of variation and degree of curvature of their data as they have no control over these measures. It is important to note that researchers must be cautious about an Exponential association as it seems to be very sensitive to the amount of variation, number of observations, and degree of curvature.

4.1 Future Research

When assessing the histogram of the estimated coefficients for the Exponential association, the histogram is unimodal and skewed. As a result, this non-normal data requires bootstrapping to transform the data. Currently, we are attempting to use bootstrap methods to bootstrap the generated data and then proceed to run a nonlinear regression. The goal of bootstrapping is to produce more consistent and reliable estimated coefficients for the Exponential performance. Following this bootstrapping methods, it is hypothesized that the Exponential performance will improve in terms of its reliability and consistency for the capture percentage and the mean of the estimated coefficients.

5 Appendix

Include the raw tables in this section.

5.1 Tables for True Value in Confidence Interval

Low Degree of Curvature

Sample Size	Variation	True Value	Exponential	Power	Logarithmic	Quadratic
n = 19 (0.5)	10%	True Value	90.30	95.10	95.50	94.60
	20%	True Value	81.90	95.80	95.90	94.90
	30%	True Value	80.00	92.80	94.70	94.60
n = 37 (0.25)	10%	True Value	91.20	95.50	93.70	95.40
	20%	True Value	85.00	95.40	93.90	94.90
	30%	True Value	78.70	94.10	94.70	95.70
n = 91 (0.1)	10%	True Value	93.20	95.60	95.30	94.40
	20%	True Value	87.30	94.20	94.50	94.40
	30%	True Value	85.30	93.50	94.30	95.70

Figure 10: Table of capture percentage of the true value for the low degree of curvature.

Medium Degree of Curvature

Sample Size	Variation	True Value	Exponential	Power	Logarithmic	Quadratic
n = 19 (0.5)	10%	True Value	93.80	94.30	96.10	94.30
	20%	True Value	90.10	93.60	94.90	94.40
	30%	True Value	85.20	91.70	95.40	95.80
n = 37 (0.25)	10%	True Value	94.40	93.50	95.70	94.70
	20%	True Value	90.90	93.90	95.60	94.10
	30%	True Value	89.30	92.60	94.60	93.80
n = 91 (0.1)	10%	True Value	94.80	94.30	94.10	94.60
	20%	True Value	92.70	94.80	95.20	95.40
	30%	True Value	90.70	93.70	94.10	95.40

Figure 11: Table of capture percentage of the true value for the medium degree of curvature.

High Degree of Curvature

Sample Size	Variation	True Value	Exponential	Power	Logarithmic	Quadratic
n = 19 (0.5)	10%	True Value	94.00	94.90	94.90	95.00
	20%	True Value	91.60	92.20	94.50	94.70
	30%	True Value	89.60	92.00	94.40	95.90
n = 37 (0.25)	10%	True Value	93.80	94.20	95.50	95.30
	20%	True Value	92.70	94.50	95.80	95.10
	30%	True Value	90.40	90.80	94.60	95.20
n = 91 (0.1)	10%	True Value	94.70	94.20	95.10	94.80
	20%	True Value	93.50	93.60	95.40	94.80
	30%	True Value	92.60	93.80	96.00	94.60

Figure 12: Table of capture percentage of the true value for the high degree of curvature.

5.2 Tables for Mean of the Estimated Coefficients

Low Degree of Curvature

Sample Size	Variation	Mean/SD	Exponential	Power	Logarithmic	Quadratic
n = 19 (0.5)	10%	Mean	15.81	2.02	20.12	12.10
		SD	10.11	0.24	1.63	4.23
	20%	Mean	19.44	2.05	19.96	11.71
		SD	16.47	0.47	3.28	8.57
	30%	Mean	22.70	2.14	19.73	11.82
		SD	20.68	0.84	4.90	13.30
n = 37 (0.25)	10%	Mean	14.19	2.00	19.93	11.99
		SD	7.51	0.17	1.29	3.16
	20%	Mean	18.11	2.04	19.98	12.20
		SD	15.48	0.32	2.52	6.49
	30%	Mean	20.20	2.04	20.14	11.99
		SD	19.83	0.54	3.71	9.57
n = 91 (0.1)	10%	Mean	13.06	2.01	19.98	12.02
		SD	4.91	0.11	0.79	2.19
	20%	Mean	15.35	2.00	19.92	11.93
		SD	10.75	0.22	1.65	4.39
	30%	Mean	18.97	2.02	19.98	11.85
		SD	16.89	0.34	2.49	6.48

Figure 13: Table of mean and standard deviation of estimated coefficients for the low degree of curvature.

Medium Degree of Curvature

Sample Size	Variation	Mean/SD	Exponential	Power	Logarithmic	Quadratic
n = 19 (0.5)	10%	Mean	3.15	3.04	2.99	3.00
		SD	0.71	0.40	0.24	1.08
	20%	Mean	3.71	3.11	2.97	3.07
		SD	2.41	0.81	0.49	2.20
	30%	Mean	4.51	3.20	3.03	3.01
		SD	4.24	1.39	0.75	3.18
n = 37 (0.25)	10%	Mean	3.09	3.02	2.99	3.02
		SD	0.48	0.29	0.18	0.82
	20%	Mean	3.37	3.04	2.99	2.97
		SD	1.47	0.57	0.36	1.71
	30%	Mean	3.85	3.16	3.02	2.97
		SD	2.74	0.99	0.56	2.47
n = 91 (0.1)	10%	Mean	3.02	3.00	2.99	3.00
		SD	0.29	0.18	0.12	0.53
	20%	Mean	3.16	3.01	3.01	3.02
		SD	0.73	0.37	0.24	1.07
	30%	Mean	3.27	3.06	2.98	2.96
		SD	1.29	0.59	0.39	1.59

Figure 14: Table of mean and standard deviation of estimated coefficients for the medium degree of curvature.

High Degree of Curvature

Sample Size	Variation	Mean/SD	Exponential	Power	Logarithmic	Quadratic
n = 19 (0.5)	10%	Mean	2.03	12.14	0.99	9.81
		SD	0.22	2.30	0.08	3.45
	20%	Mean	2.16	12.67	1.00	10.03
		SD	0.77	4.25	0.17	6.95
	30%	Mean	2.40	12.73	0.99	10.09
		SD	1.38	5.38	0.25	10.08
n = 37 (0.25)	10%	Mean	2.02	12.18	0.99	9.82
		SD	0.17	1.79	0.06	2.53
	20%	Mean	2.05	12.49	1.00	10.36
		SD	0.34	3.32	0.12	5.16
	30%	Mean	2.23	12.27	0.99	9.78
		SD	1.04	4.53	0.19	7.96
n = 91 (0.1)	10%	Mean	2.00	12.14	1.00	9.99
		SD	0.10	1.21	0.04	1.68
	20%	Mean	2.01	12.24	1.00	10.02
		SD	0.21	2.51	0.08	3.38
	30%	Mean	2.06	12.36	0.99	9.98
		SD	0.35	3.56	0.12	5.11

Figure 15: Table of mean and standard deviation of estimated coefficients for the high degree of curvature.

5.3 Linear vs Nonlinear Regression

As mentioned above, this paper uses nonlinear regression techniques to estimate the coefficients; however, it is common for some disciplines to use linearization to transform the nonlinear associations into a linear association. Afterwards, the research uses linear regression to estimate the coefficients. This paper argues that researchers should not transform their nonlinear data and estimate the coefficients by using nonlinear techniques as described above. In Figure 4, there are two different nonlinear association models used to visual the difference between the estimation techniques. The blue line refers to a classic linearization transformation and then using a linear regression to estimate the line, whereas the red line is simply a nonlinear regression to estimate the line. It is evidence that there is a clear distinction between the two lines and the red line estimates the best fits to the data when compared to the blue line.

Linearization v. Nonlinear Regression

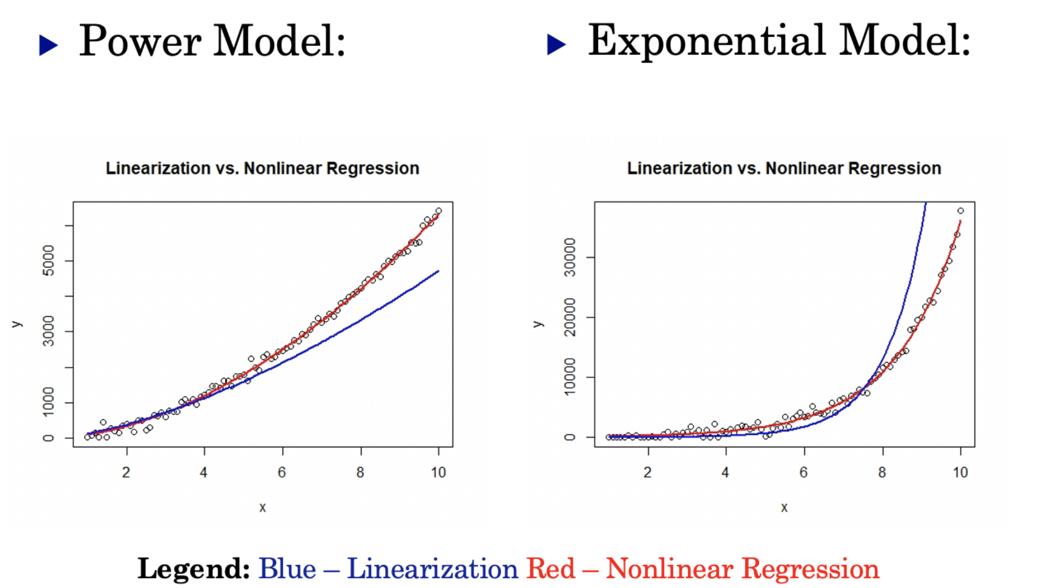

Figure 16: This figure shows the difference in the estimation techniques when using Linearization and Nonlinear Regression.

6 Acknowledgments

I would like to thank Dr. Hydorn for her comments and assistance on this research paper. Any errors are the author's errors.

7 References

Mendenhall, W., Sincich, T. (2012). A second course in statistics: Regression analysis. Upper Saddle River, NJ: Pearson Education.

Studenmund, A. H. (2017). Using econometrics: A practical guide. Boston: Pearson.

Klairung Samart, Naratip Jansakul Mitchai Chongcheawchamnan (2018) Exact bootstrap confidence intervals for regression coefficients in small samples, Communications in Statistics - Simulation and Computation, 47:10, 2953-2959, DOI: 10.1080/03610918.2017.1364386